EECS 598: Statistical Learning Theory, Winter 2014 **Topic 7**

Dyadic Decision Trees

Lecturer: Clayton Scott *Scribe: Pin-Yu Chen, Gopal Nataraj*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

1 Introduction

These notes introduce a new kind of classifier called a dyadic decision tree (DDT). We also introduce a discrimination rule for learning a DDT that achieves the optimal rate of convergence, $\mathbb{E}R(\widehat{h}_n) - R^* = O(n^{-1/d})$, for the box-counting class, which was defined in the previous set of notes. This improves on the rate of $\mathbb{E}R(\widehat{h}_n) - R^* = O(n^{-1/(d+2)})$ for the histogram sieve estimator from the previous notes.

Dyadic decision trees are based on recursively splitting the input space at the midpoint along some dimension. This is in contrast to conventional decision trees that allow the splits to occur at any point. Yet DDTs can still approximate complex decision boundaries, and the restriction to dyadic splits makes it possible to globally optimize a complexity penalized empirical risk criterion, in contrast to mainstream methods for decision tree learning that first perform greedy growing followed by pruning. These notes will not discuss implementation of the discrimination rules, but the interested reader can find algorithms and computational considerations discussed in [1, 2, 3].

2 Recursive Dyadic Partitions

Assume $\mathcal{X} = [0,1]^d$. A *recursive dyadic partition* (RDP) is a partition of $\mathcal{X}$ obtained by applying the following two rules:

- $\{[0,1]^d\}$ is a RDP.
- If $\{A_1, \ldots, A_k\}$ is a RDP, where A_i is a rectangle, then so is $\{A_1, \ldots, A_{i-1}, A_i^1, A_i^2, A_{i+1}, \ldots, A_k\}$, where A_i^1, A_i^2 are obtained by splitting A_i at its midpoint along some dimension. Note that every $A = \prod_{\ell=1}^d [a_\ell, b_\ell]$, where a_ℓ, b_ℓ are dyadic rational numbers in the form $r/2^s$, $0 \le r \le 2^s$.

A simple illustration of a RDP is shown in Fig. 1 in the case $d = 2$.

A *dyadic decision tree* (DDT) is a classifier that is constant on a RDP. Let $\mathcal{T} = \{\text{all DDTs}\}$, and $\mathcal{T}_m = \{\text{all DDTs where all cells have side length} \ge \frac{1}{m}, \}$, where m is a power of 2. If $m = 2^J$, then J is the maximum number of splits along any dimension.

Note that a histogram partition is a special case of a recursive dyadic partition, where every cell in the partition is a hypercube of the same size. By pruning back cells that do not intersect the Bayes decision boundary, a dyadic decision tree can achieve the same approximation error as a histogram, but since there are fewer cells in the partition, we can get a tighter bound on the estimation error.

3 Uniform Deviation Bound for DDTs

We will use prefix codes to derive a uniform deviation bound (UDB) for DDTs. Let $\mathcal{C} = \{c_1, c_2, \ldots\}$ be a set of finite length binary strings. We say $\mathcal{C}$ is a prefix code iff no c_i is a prefix of another c_j. Let ℓ_i be the codeword length of c_i. The following fact from information theory will be useful:

$\exists$ a prefix code $\mathcal{C}$ with codeword lengths $\ell_i \Leftrightarrow \sum_i 2^{-\ell_i} \le 1$.

The inequality on the right-hand side is known as Kraft's inequality. We will only need the forward implication: if $\mathcal{C}$ is a prefix code, then $\sum_i 2^{-\ell_i} \le 1$.

Suppose $\mathcal{C}$ is a prefix code for $\mathcal{T}$. Let $\ell(h)$ denote the length of the codeword assigned to h.

Proposition 1. *Let $\delta > 0$. With probability $\ge 1-\delta$,*

$$\forall h \in \mathcal{T} \quad \left|\widehat{R}_n(h) - R(h)\right| \le \sqrt{\frac{\ell(h)\ln 2 + \ln(2/\delta)}{2n}}.$$

Proof. For a fixed $h \in \mathcal{T}$, by Hoeffding's inequality, for any $\delta_h > 0$,

$$\Pr\left(\left|\widehat{R}_n(h) - R(h)\right| \ge \sqrt{\frac{\ln(2/\delta_h)}{2n}}\right) \le \delta_h \ \left(\text{by setting } \delta_h = 2e^{-2n\epsilon_h^2}\right).$$

Let $\delta_h = \delta 2^{-\ell(h)}$. By the union bound,

$$\Pr\left(\exists\ h\ \left|\widehat{R}_n(h) - R(h)\right| \ge \sqrt{\frac{\ell(h)\ln 2 + \ln(2/\delta)}{2n}}\right) \le \sum_{h\in\mathcal{T}} \delta 2^{-\ell(h)} \le \delta \text{ (by Kraft's inequality).}$$

□

Note that the above argument holds for any countable set $\mathcal{H}$ of classifiers. When $\mathcal{H}$ is finite, we can take $\ell(h) = \log_2|\mathcal{H}|$ for all $h \in \mathcal{H}$, in which case we recover the UDB for finite $\mathcal{H}$ derived previously.

Now let's determine a prefix code for $\mathcal{T}$. Let $k = |h| :=$ number of leaf nodes in DDT, i.e., the number of cells in the associated RDP. The total number of nodes is $2k-1$ (the number of internal nodes is $k-1$). A simple illustration is shown in Fig. 2.

- To encode the tree structure, we use $2k-1$ bits. The encoding procedure is implemented as follows. Staring at the root node, scan through the nodes from left to right and then top to bottom. If a node is split, assign a 1, otherwise assign a 0. It is easy to verify that by construction this code is a prefix code for the tree structure. An illustration is shown in Fig. 3.
- To encode the dimension being split at each internal node, we append $(k-1)\log_2 d$ bits to the prefix code for tree structure.
- To encode the class labels of the leaf nodes, we append k bits to the prefix code for tree structure and splitting dimensions.

Summing up, $\ell(h) = (3k-1) + (k-1)\log_2 d \le (3+\log_2 d)k = (3+\log_2 d)|h|$. Denote $\kappa = (3+\log_2 d)\ln 2$.

Corollary 1. *With probability $\ge 1-\delta$,*

$$\forall h \in \mathcal{T}, \quad \left|\widehat{R}_n(h) - R(h)\right| \le \sqrt{\frac{\kappa|h| + \ln(2/\delta)}{2n}}.$$

4 Convergence Rates of Dyadic Decision Trees

The above bound motivates the following discrimination rule:

$$\widehat{h}_n = \underset{h\in\mathcal{T}_m}{\arg\min}\, \widehat{R}_n(h) + \Phi_n(h) \tag{1}$$

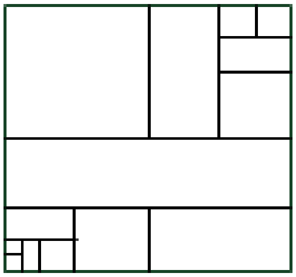


Figure 1: Illustration of recursive dyadic partition (RDP) for $d = 2$. The bounding box is $\mathcal{X} = [0,1]^d$.

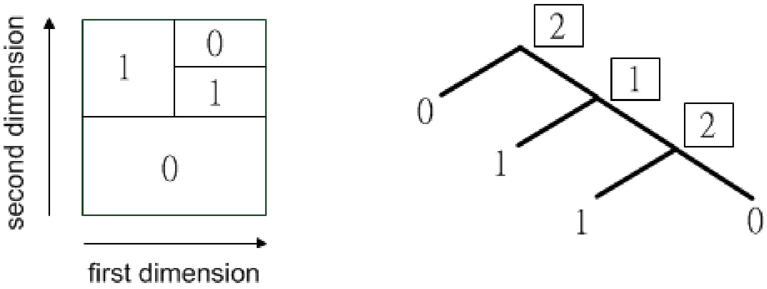


Figure 2: Illustration of encoding dyadic decision tree (DDT) structure using prefix code. The number i in the box indicates partition on ith dimention.

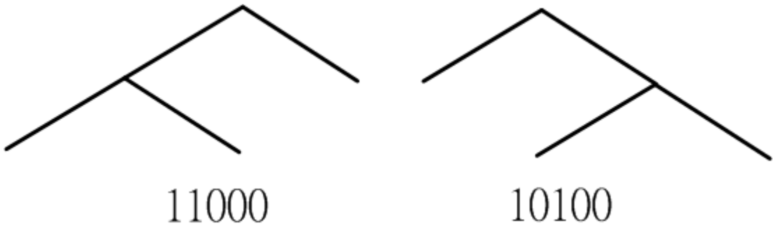
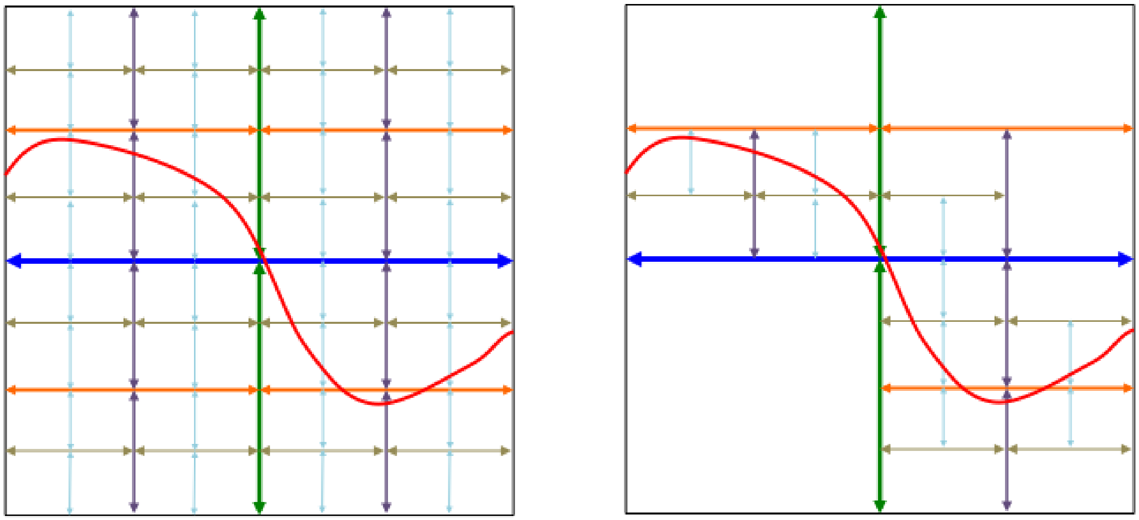


Figure 3: Illustration of encoding tree structure



(a) h_m^0 (b) h_m^*

Figure 4: Examples of recursive dyadic partitions (RDPs) of (a) a cyclic DDT h_m^0 and (b) the corresponding pruned classifier h_m^* for a certain Bayes decision boundary (red), with dimension $d = 2$ and depth $m = 6$.

where

$$\Phi_n(h) := \sqrt{\frac{\kappa|h| + \ln(2/\delta)}{2n}}\,.$$

This optimization problem can be interpreted as a form of penalized empirical risk minimization, where $\Phi_n(h)$ quantifies the complexity of h. $\widehat{h}_n$ therefore achieves a balance between data fit and model complexity.

As of now, the depth m is a free parameter. We will show that appropriate selection of m will allow $\widehat{h}_n$ to achieve better convergence rates than the histogram sieve estimator.

Proposition 2. *With probability at least* $1 - \delta$*,*

$$R(\widehat{h}_n) - R^* \leq \inf_{h \in \mathcal{T}_m} \left\{ R(h) - R^* + 2\Phi_n(h) \right\}. \tag{2}$$

Proof. Applying the bound of Corollay 1 twice, we obtain *w.p.* $\geq 1 - \delta$, $\forall h \in \mathcal{T}_m$,

$$\begin{aligned} R(\widehat{h}_n) &\leq \widehat{R}_n(\widehat{h}_n) + \Phi_n(\widehat{h}_n) \\ &\leq \widehat{R}_n(h) + \Phi_n(h) \\ &\leq R(h) + 2\Phi_n(h)\,, \end{aligned} \tag{3}$$

where in the second inequality, we use the definition of $\widehat{h}_n$. As h is arbitrary, we can select h to come arbitrarily close to the infimum. Subtracting R^* from both sides gives the desired result. □

The following definition is used in our proofs of rates of convergence.

Definition 1. *A DDT is* cyclic *if the dimensions along which its splits are taken in a cyclic order.*

Theorem 1. *Suppose* $P_{XY} \in \mathcal{B}$*, where* $\mathcal{B}$ *denotes the box-counting class. As* $n \to \infty$*, allow* m *to increase as* $m \sim n^{\frac{1}{d+1}}$*. Then* $\widehat{h}_n$ *defined as in* (1) *satisfies*

$$\mathbb{E}R(\widehat{h}_n) - R^* = O(n^{-\frac{1}{d+1}})\,. \tag{4}$$

Proof. We find a *particular* DDT classifier h_m^* whose approximation and estimation errors achieve the claimed convergence rate. Then the infimum will achieve the rate as well.

Let $h_m^0 \in \mathcal{T}_m$ be a cyclic DDT where every leaf node is a cube with side length $\frac{1}{m}$. Then every leaf node is at maximum depth $d \log_2 m$. Assume labels are assigned to minimize $R(h_m^0)$. Let h_m^* be obtained by "pruning" alls cells of h_m^0 whose parents do not intersect the Bayes decision boundary (BDB), i.e., all cells such that neither the cell nor its sibling intersect the BDB. Examples of an h_m^0 and the corresponding h_m^* for a specific Bayes decision boundary are given in Fig. 4.

Observe that although h_m^* has significantly fewer splits than h_m^0, it suffers no loss in resolution as compared to h_m^0 around the Bayes decision boundary. Indeed,

$$R(h_m^*) - R^* = R(h_m^0) - R^*,$$

and by the same argument as for a histogram classifier, we have

$$R(h_m^0) - R^* = O\left(\frac{1}{m}\right).$$

Furthermore, while h_m^0 has m^d leaf nodes, we can show that h_m^* has $O(m^{d-1})$ leaf nodes. We state this result, and a useful intermediate result, in the following lemma.

Lemma 1. *The number of nodes in h_m^* at depth j, including internal nodes, that intersect the Bayes decision boundary, is that most $C2^{\lceil j/d \rceil}(d-1)$. Furthermore, $|h_m^*| \le 4dCm^{d-1}$ where C is the constant from condition* **(B)** *in the definition of the box-counting class.*

Proof. Write $j = (p-1)d + q$ where $1 \le p \le \log_2 m$ and $1 \le q \le d$. Let N_j denote the number of nodes at depth j in h_m^* intersecting the BDB. Clearly, if a node at depth j intersects the BDB, then it contains a descendent at depth pd that also intersects the BDB, and therefore $N_j \le N_{pd}$. Note that all nodes at depth pd are hypercubes with side length 2^{-p}. By the box-counting assumption,

$$N_{pd} \le C(2^p)^{d-1} = C2^{\lceil j/d \rceil (d-1)}.$$

This establishes the first part of the lemma. Applying this result, the total number of nodes of h_m^* at any depth that intersect the BDB is at most

$$\sum_{j=1}^{d \log_2 m} C2^{\lceil j/d \rceil (d-1)} \le \sum_{p=1}^{\log_2 m} dC2^{p(d-1)} \le 2dC2^{(d-1)\log_2 m} = 2dCm^{d-1}.$$

Now, to establish the second part of the lemma, notice that every leaf node of h_m^* either intersects the Bayes decision boundary, or its sibling intersects the Bayes decision boundary. Therefore $|h_m^*| \le 4dCm^{d-1}$. □

Take Ω to be the event in Prop. 2 that holds with high probability. For $\delta = \frac{1}{n}$, we have by the law of total expectation,

$$\begin{aligned}
\mathbb{E}R(\widehat{h}_n) - R^* &= \underbrace{\Pr(\Omega)}_{\le 1} \underbrace{\mathbb{E}\Big[R(\widehat{h}_n) - R^* | \Omega\Big]}_{\le R(h_m^*) - R^* + 2\Phi_n(h_m^*)} + \underbrace{\Pr(\Omega^C)}_{\le \frac{1}{n}} \underbrace{\mathbb{E}\Big[R(\widehat{h}_n) - R^* | \Omega^C\Big]}_{\le 1} \\
&\le R(h_m^*) - R^* + 2\Phi_n(h_m^*) + \frac{1}{n}. \\
&= O\Big(\frac{1}{m} + \sqrt{\frac{1}{n}(m^{d-1} + \ln n)} + \frac{1}{n}\Big) \\
&= O\Big(\frac{1}{m} + \sqrt{\frac{m^{d-1}}{n}}\Big). \qquad (5)
\end{aligned}$$

If m grows as $m \sim n^{\frac{1}{d+1}}$, then both terms in the last expression decay as $O(n^{-\frac{1}{d+1}})$, completing the proof. □

5 A Spatially Adaptive Penalty

The previous result suggests that a penalty based only on tree size is not sufficient for attaining the optimal rate of convergence. We now develop an alternative penalty that leads to the optimal rate.

Recall that every $h \in \mathcal{T}$ is associated with a RDP of $\mathcal{X} = [0,1]^d$. Let us denote the partition associated to h by $\Pi(h) = \{A_1, \ldots, A_k\}$. These are just the leaf nodes of h. Note that Π is a many-to-one mapping: different DDTs can have the same RDP. Observe that

$$R(h) - \widehat{R}_n(h) = \sum_{A \in \Pi(h)} R(h,A) - \widehat{R}_n(h,A) \tag{6}$$

where

$$R(h,A) := P_{XY}\big(\{h(X) \neq Y\} \cap \{X \in A\}\big)\,, \text{ and}$$
$$\widehat{R}_n(h,A) := \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{\{h(X_i)\neq Y_i\}\cap\{X_i \in A\}\}}\,.$$

Because $n\widehat{R}_n(h,A) \sim \text{binom}(n, R(h,A))$, we could use Hoeffding's inequality to obtain a convergence rate, but this will not lead to the desired rate. We instead use the relative Chernoff bound:

Lemma 2 (Relative Chernoff Bound). *Let $Z_1, \ldots, Z_n \overset{i.i.d.}{\sim} \text{Ber}(p)$ and $\widehat{p} = \frac{1}{n}\sum_{i=1}^{n} Z_i$. Then $\forall \epsilon \in [0,1]$,*

$$\Pr\big(\widehat{p} \le (1-\epsilon)p\big) \le e^{-np\epsilon^2/2}\,. \tag{7}$$

Equivalently, taking $\delta := e^{-np\epsilon^2/2}$ and thus $\epsilon = \sqrt{\frac{2\ln(1/\delta)}{np}}$, we have that with probability at least $1-\delta$,

$$p \le \widehat{p} + \sqrt{\frac{2p\ln(1/\delta)}{n}}\,. \tag{8}$$

Proof. Refer to [5]. □

By combining the relative Chernoff bound with the decomposition in (6), we can arrive at the following uniform deviation bound. To state the next result, let $\mathcal{A}$ be the set of all cells that belong to some RDP, let $\ell(A)$ be the length of a codeword for A in a prefix code for $\mathcal{A}$, and denote $p_A := P_X(A)$ for any $A \in \mathcal{A}$.

Proposition 3. *With probability at least $1 - \frac{1}{n}$, $\forall h \in \mathcal{T}$,*

$$|R(h) - \widehat{R}_n(h)| \le \sum_{A \in \Pi(h)} \sqrt{\frac{2p_A[\ell(A)\ln 2 + \ln n]}{n}}\,. \tag{9}$$

Proof. By the Relative Chernoff Bound, we know that for each $A \in \mathcal{A}$, with probability at least $1 - \delta_A$,

$$R(h,A) - \widehat{R}_n(h,A) \le \sqrt{\frac{2R(h,A)\ln(1/\delta_A)}{n}}. \tag{10}$$

Taking $\delta_A := \frac{1}{n}2^{-\ell(A)}$, we know by Kraft's inequality that $\sum_{A\in\mathcal{A}} \delta_A \le \frac{1}{n}$. Thus, by the union bound, and noting that $R(h,A) \le p_A$, we have that with probability at least $1 - \frac{1}{n}$, $\forall h \in \mathcal{T}$,

$$\begin{aligned}
R(h) - \widehat{R}_n(h) &= \sum_{A\in\Pi(h)} R(h,A) - \widehat{R}_n(h,A) \\
&\le \sum_{A\in\Pi(h)} \sqrt{\frac{2R(h,A)[\ell(A)\ln 2 + \ln n]}{n}} \\
&\le \sum_{A\in\Pi(h)} \sqrt{\frac{2p_A[\ell(A)\ln 2 + \ln n]}{n}}
\end{aligned}$$

To establish the absolute value in the bound, consider the complementary classifier $h^C(x) = 1 - h(x)$. Then on the same event that the previous bound holds on,

$$\begin{aligned}\widehat{R}_n(h) - R(h) &= \left[1 - \widehat{R}_n(h^C)\right] - \left[1 - R(h^C)\right] = R(h^C) - \widehat{R}_n(h^C) \\ &\leq \sum_{A \in \Pi(h^C)} \sqrt{\frac{2p_A[\ell(A)\ln 2 + \ln n]}{n}} \\ &\leq \sum_{A \in \Pi(h)} \sqrt{\frac{2p_A[\ell(A)\ln 2 + \ln n]}{n}}. \end{aligned} \tag{11}$$

Note that the last line follows because $\Pi(h)$ and $\Pi(h^C)$ are the same partition. □

The result in Proposition 3 does not quite yet give us a useful penalty:

1. In practice, p_A is unknown. For the box-counting class, P_X has a density f such that $\forall x,\ f(x) \leq B$, where B is a constant. We will assume B is known. Now the volume $\lambda(A)$ of a cell at depth j is just 2^{-j}. Thus, p_A can be bounded as

$$p_A = P_X(A) = \int_A f(x)\,\mathrm{d}\,x \leq B\lambda(A) = B2^{-j(A)}\,, \tag{12}$$

   where $j(A)$ denotes the depth of A. (It is not necessary to assume B is known. One can upper bound p_A by its empirical counterpart to obtain a data-dependent penalty, and the following analysis carries through in a similar way. The details are relatively straightforward, but are omitted in the interest of brevity. The interested reader may refer to [2].)

2. We also need to design a prefix code for $\mathcal{A}$. The following code suffices: Use

   - $j + 1$ bits to encode the depth of A: j 0s followed by a 1;
   - $j \log_2 d$ bits to encode the dimension along with splits are taken; and
   - j bits to encode whether the ancestors of A split “left” or “right.”

   This scheme produces codewords of length $\ell(A) = (2j + 1) + j\log_2 d \leq (3 + \log_2 d)j$. Denote $\kappa = (3 + \log_2 d)\ln 2$ as before.

We can combine these bounds with Proposition 3 to finally conclude:

Corollary 2. *With probability at least* $1 - \frac{1}{n}$,

$$|R(h) - \widehat{R}_n(h)| \leq \sum_{A \in \Pi(h)} \sqrt{\frac{2B2^{-j(A)}[\kappa j(A) + \ln n]}{n}} =: \Phi'_n(h)\,. \tag{13}$$

We now define a new discrimination rule based on the above penalty:

$$\widehat{h}_n = \arg\min_{h \in \mathcal{T}_m} \widehat{R}_n(h) + \Phi'_n(h)\,. \tag{14}$$

As before, we have the following performance guarantee.

Proposition 4. *With probability at least* $1 - \frac{1}{n}$*, the rule in* (14) *satisfies*

$$R(\widehat{h}_n) - R^* \leq \inf_{h \in \mathcal{T}_m} \left\{R(h) - R^* + 2\Phi'_n(h)\right\}. \tag{15}$$

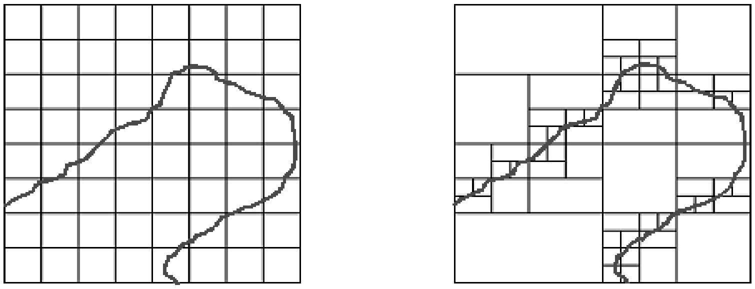


Figure 5: The penality Φ_n penalizes both trees above the same, whereas Φ'_n favors the partition on the right.

Proof. The argument follows that of Proposition 2, replacing $\Phi_n(h)$ with the new $\Phi'_n(h)$ penalty. □

Observe that this new penalty $\Phi'_n(h)$ has a different structure compared to the previous penalty $\Phi_n(h)$. Whereas $\Phi_n(h)$ depended on h only through $|h|$, the new penalty depends also on the depth (equivalently, volume) of the cells. While Φ_n will not distinguish between the two trees shown in Figure 5, the new penalty Φ'_n will prefer the tree on the right. More generally, the new penalty prefers unbalanced trees to balanced trees. Since unbalanced trees are sufficient for accurately approximating decision boundaries in the box-counting class, the spatially adaptive penalty provides a tighter bound on the estimation error for the same approximation error. This intuition is made precise in the following, the main result of this section.

Theorem 2. *Suppose $P_{XY} \in \mathcal{B}$, where $\mathcal{B}$ denotes the box-counting class. As $n \to \infty$, allow m to increase as $m \sim (n/\log n)^{1/d}$. The discrimination rule $\widehat{h}_n$ in* (14) *satisfies*

$$\mathbb{E}R(\widehat{h}_n) - R^* = O\left(\left(\frac{\log n}{n} \right)^{\frac{1}{d}} \right). \tag{16}$$

Proof. Let h^*_m be as in the proof of Thm. 1. It suffices to show that the approximation and estimation errors corresponding to h^*_m achieve the claimed convergence rate. Then the infimum must also achieve this rate.

We previously argued that the approximation error $R(h^*_m) - R^* = O(\frac{1}{m})$. When $m \sim \left(\frac{n}{\log n}\right)^{\frac{1}{d}}$, this becomes $O\left(\left(\frac{\log n}{n}\right)^{\frac{1}{d}} \right)$. This part of the argument is unchanged except for the rate at which m grows. Henceforth we focus on bounding $\Phi'_n(h^*_m)$.

Observe that because $j(A) \leq d \log_2 m = O(\log n)$,

$$\Phi'_n(h^*_m) = O\left(\sum_{A \in \Pi(h^*_m)} \sqrt{2^{-j(A)} \frac{\log n}{n}} \right) = O\left(\sqrt{\frac{\log n}{n}} \sum_{A \in \Pi(h^*_m)} \sqrt{2^{-j(A)}} \right). \tag{17}$$

To bound the interior summation, note that there exist unique $p \in \{1, \ldots, \log_2(m)\}$ and $q \in \{1, \ldots, d\}$

such that $j(A) = (p-1)d + q$. Let $\Pi_{p,q}(h) = \{A \in \Pi(h) \,|\, j(A) = (p-1)d + q\}$. Then,

$$
\begin{aligned}
\Phi'_n(h^*_m) &= O\Bigg(\sqrt{\frac{\log n}{n}} \sum_{A \in \Pi(h^*_m)} \sqrt{2^{-j(A)}}\Bigg) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} \sum_{p=1}^{\log_2 m} \sum_{q=1}^{d} \sum_{A \in \Pi_{p,q}(h^*_m)} \sqrt{2^{-(p-1)d-q}}\Bigg) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} \sum_{p=1}^{\log_2 m} \sum_{q=1}^{d} C 2^{p(d-1)} \sqrt{2^{-(p-1)d-q}}\Bigg) \qquad (18) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} \sum_{p=1}^{\log_2 m} C d 2^{p(d-1)} \sqrt{2^{-(p-1)d}}\Bigg) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} \sum_{p=1}^{\log_2 m} 2^{p(\frac{d}{2}-1)}\Bigg) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} 2^{(\frac{d}{2}-1)\log_2 m}\Bigg) \\
&= O\Bigg(\sqrt{\frac{\log n}{n}} m^{(\frac{d}{2}-1)}\Bigg) \\
&= O\Bigg(\left[\frac{\log n}{n}\right]^{\frac{1}{d}}\Bigg).
\end{aligned}
$$

Eqn. (18) follows from Lemma 1. The final line follows by allowing m to grow as $m \sim \left(\frac{n}{\log n}\right)^{\frac{1}{d}}$. Thus, both the approximation and estimation errors are of order $O\big(\left(\frac{\log n}{n}\right)^{\frac{1}{d}}\big)$, completing the proof. □

Exercises

1. We have not yet leveraged the full flexibility of DDTs (we will do so in the next lecture).

   (a) Let $\mathcal{T}^{\text{cyc}} \subseteq \mathcal{T}$ denote the set of all cyclic dyadic decision trees. Design a prefix code for $\mathcal{T}^{\text{cyc}}$ that is more concise than the one we designed for $\mathcal{T}$. State an analogue to Corollary 1 for cyclic DDTs, and briefly explain why penalized empirical risk minimization over $\mathcal{T}^{\text{cyc}}_m := \mathcal{T}^{\text{cyc}} \cap \mathcal{T}_m$ can also achieve the rate of convergence in Theorem 1.

   (b) Let $\mathcal{A}^{\text{cyc}} \subseteq \mathcal{A}$ denote the set of all cells in partitions associated with cyclic dyadic decision trees. Design a prefix code for $\mathcal{A}^{\text{cyc}}$ that is more concise than the one we designed for $\mathcal{A}$. State an analogue to Corollary 2 for cyclic DDTs, and briefly explain why penalized empirical risk minimization over $\mathcal{T}^{\text{cyc}}_m := \mathcal{T}^{\text{cyc}} \cap \mathcal{T}_m$ can also achieve the rate of convergence in Theorem 2.

2. We have not yet harnessed the full power of penalized empirical risk minimization as an algorithm for learning DDTs (we will do so in the next lecture). In particular, the rates of convergence in Theorems 1 and 2 can be obtained with sieve estimators. Thus, define $\mathcal{T}_{m,k} = \{h \in \mathcal{T}_m : |h| \leq k\}$. Let us view $m = m(n)$ and $k = k(n)$, and define the sieve estimator

$$\widehat{h}_n = \underset{h \in \mathcal{T}_{m(n),k(n)}}{\arg\min}\ \widehat{R}_n(h).$$

   For simplicity, assume the empirical risk minimizer exists. Give sufficient conditions on $m(n)$ and $k(n)$ such that the above sieve estimator achieves the rate of convergence in Theorem 2. *Hint:* It's not necessary to do any additional analysis. Just combine properties of sieve estimators with the analysis in these notes.

References

[1] C. Scott, “Tree pruning with subadditive penalties,” *IEEE Trans. Signal Processing*, vol. 53, no. 12, pp. 4518-4525, 2005.

[2] C. Scott and R. Nowak, “Minimax-Optimal Classification with Dyadic Decision Trees,” *IEEE Trans. Inform. Theory*, vol. 52, pp. 1335-1353, 2006.

[3] G. Blanchard, C. Schäfer, Y. Rozenholc, K-R. Müller, “Optimal Dyadic Decision Trees,” *Machine Learning*, vol. 66, nos. 2-3, pp. 209-242, 2007.

[4] Thomas M. Cover and Joy A. Thomas, *Elements of Information Theory*, Wiley-Interscience, 1991.

[5] Torben Hagerup and Christine Rüb. “A Guided Tour of Chernoff Bounds,” *Inform. Proc. Letters*, vol. 33, pp. 305-308, 1990.